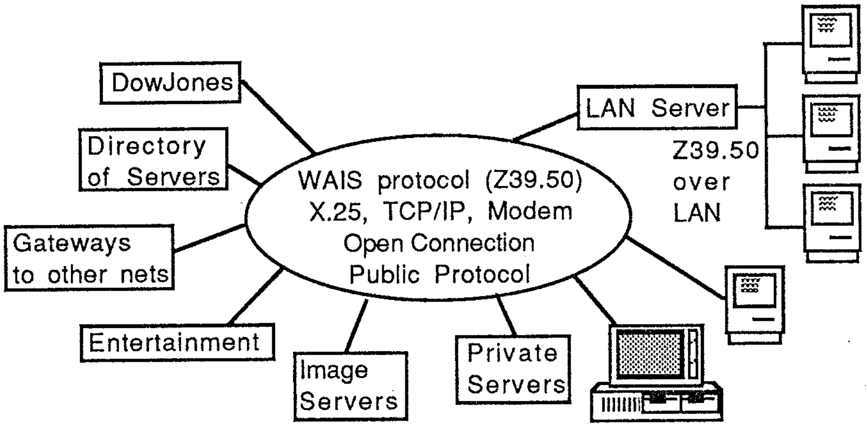
Thinking Machines Corporation

Wide Area Information Servers

Brewster Kahle

December 1991



Users Needs:
- Selecting Servers
- Answering Questions
- Organizing Responses

Architecture Issues:
- Scalability
- Security
- Business model for servers
- Reliable Access

The Wide Area Information Servers (WAIS) system is an electronic publishing system that helps end users find unstructured information located on remote machines. It is composed of user interfaces, available for most machines, and server software. Started by Thinking Machines, this system is becoming a standard for information distribution in the internet environment. Since many components are available for free, please try the system!

What does WAIS do? Users on different platforms can access personal, company, and published information from one interface. The information can be anything: text, pictures, voice, or formatted documents. Since a single computer-to-computer protocol is used, information can be stored anywhere on different types of machines. Anyone can use this system since it uses natural language questions to find relevant documents. Relevant documents can be fed back to a server to refine the search. This avoids complicated query languages and vendor specific systems. Successful searches can be automatically run to alert the user when new information becomes available.

How does WAIS work? The servers take a users question and do their best to find relevant documents. The servers, at this point, do not "understand" the users English language question, rather they try to find documents that contain those words and phrases and ranks then based on heuristics. The user interfaces (clients) talk to the servers using an extension to a standard protocol Z39.50. Using a public standard allows vendors to compete with each other, while bypassing the usual proprietary protocol period that slows development. Thinking Machines is giving away an implementation of this standard to help vendors develop clients and servers.

What WAIS servers exist? Even though the system is very new, there are already over 100 servers on the internet. Over 5000 people have used WAIS in 20 countries.

- Thinking Machines operates a Connection Machine on the internet for free use. The databases it supports are some patents, a collection of molecular biology abstracts, a cookbook, and the *CIA World Factbook*.
- MIT supports a poetry server with a great deal of classical and modern poetry. Cosmic is serving descriptions of government software packages. The Library of Congress has plans to make their catalog available on the protocol.
- Weather maps and forecasts are made available by Thinking Machines as a repackaging of existing information.
- The "directory of servers" facility is operated by Thinking Machines so that new servers can be easily registered as either for-pay or for-free servers and users can find out about these services.
- Dow Jones is putting a server on their own DowVision network. This server contains the *Wall Street Journal, Barrons*, and 450 magazines. This is a for-pay server.

How can I find out more about WAIS?

- You can try a simple interface by telneting to quake.think.com, login wais.
- FTP the free software from think.com in the /wais directory.
- FTP a bibliography: /pub/wais/wais-discussion/bibliography.txt@quake.think.com
- Contact Barbara Lincoln (barbara@think.com) for more information, or Brewster Kahle the project leader.
- Subscribe to a biweekly mailing list on electronic publishing issues, and new releases; to subscribe send email to wais-discussion-request@think.com.

Brewster Kahle
Project Leader
Wide Area Information Servers
Brewster@Think.com

Business Day

The New York Times

For Shakespeare, Just Log On

Large PC Libraries Are Being Developed

By JOHN MARKOFF

The development of a nationwide data network will allow personal computer users to tap sources as large as the Library of Congress or receive their own personalized electronic newspapers.

Several innovations, taken together, have already demonstrated that searching vast computer data bases can be easier than consulting a card catalogue, and not nearly as difficult or expensive as computer searches are today. Computer users might read some Dickens more readily than they could check out David Copperfield from the local library.

Those in the industry say that users with little computer skills will soon be able to search through several terabytes of information, or several trillion characters of text, in seconds. The Library of Congress, with 80 million items, contains an estimated 25 terabytes of information.

Already, an experimental computer library has linked 150 universities to 40 sources of information, ranging from National Institutes of Health data to corporate documents and Shakespeare's plays. New software allows users to browse or zero in on particular information.

As methods of retrieving information are standardized and perfected, industry executives and computer scientists say, thousands of new services, ranging from electronic newspapers to the computer equivalent of free public libraries, will blossom. "Everyone is realizing how important it is to get into the mass market for information," said Thomas Koulopoulos, president of Delphi Consulting Group, a Boston market research firm.

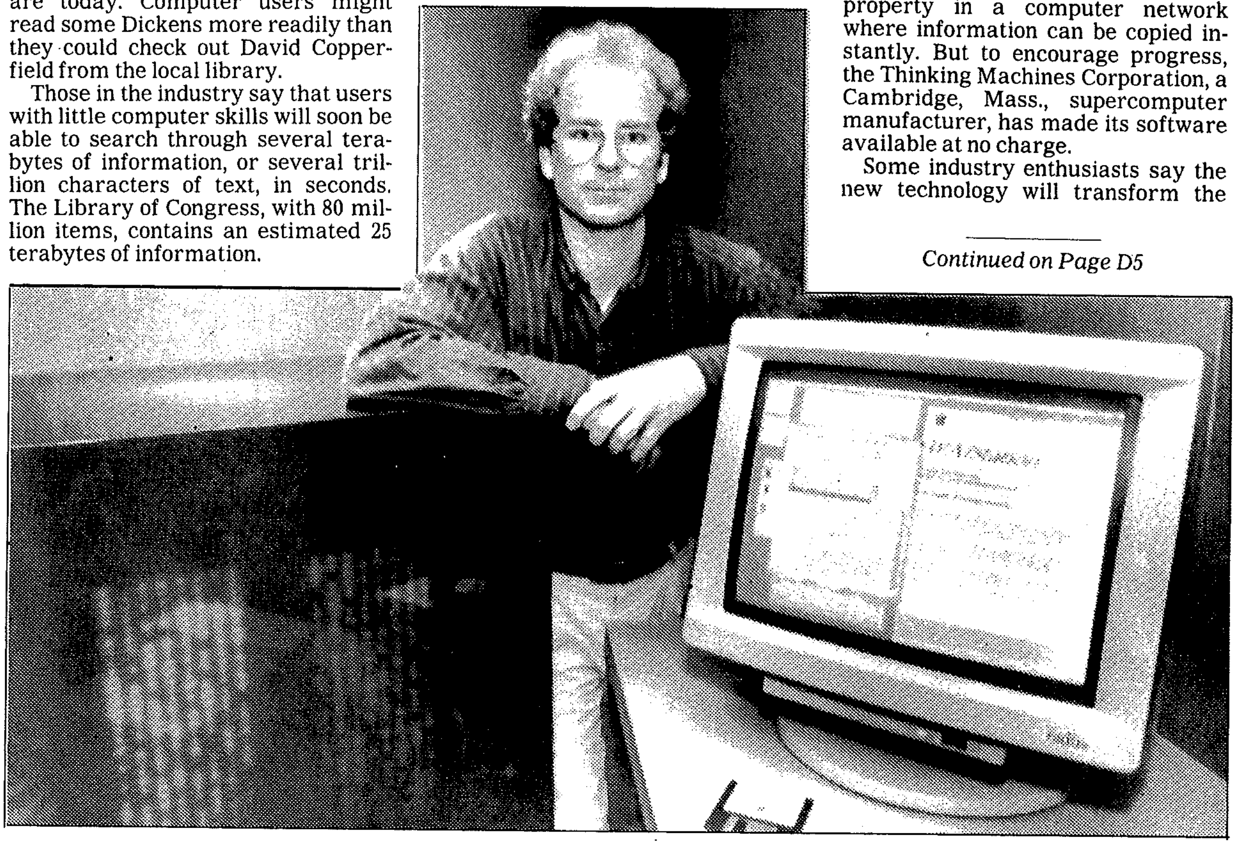
Such ready access to huge amounts of computerized information has been the dream of many in the industry. But a lack of computing power, effective software and high-speed digital networks has stalled progess until recently.

If many of the technical problems are being solved, major business and political disputes remain. The researchers acknowledge that they must resolve several questions of privacy and pricing before they can put the new methods to commercial use.

Many sources of information, like government documents, might be available free, but other services, including electronic newspapers, will be available only to those who pay. The industry has yet to settle on ways to protect and charge for intellectual property in a computer network where information can be copied instantly. But to encourage progress, the Thinking Machines Corporation, a Cambridge, Mass., supercomputer manufacturer, has made its software available at no charge.

Some industry enthusiasts say the new technology will transform the

Continued on Page D5



Mike Theiler for The New York Times

Brewster Kahle was the leader of the development team at the Thinking Machines Corporation for a nationwide computerized library system. His team's software links a CM2A Connection Machine, left, with a personal computer or work station like the Apple Macintosh II at right. Using high-speed data highways, the two machines can function together although they may be thousands of miles apart.

For Shakespeare, Just Log On

Continued From First Business Page

way computerized information is sold. Mitchell Kapor, the founder of the Lotus Development Corporation, predicts the growth of a new industry as significant as the personal computer business. Some companies, like Dow Jones & Company, that already provide computerized information over telephone lines have taken part in developing the new computer library.

The Search Is Simplified

In 1989, Thinking Machines enlisted the support of Dow Jones, Apple Computer Inc. and the KPMG Peat Marwick accounting and consulting firm to design the computer library, called Wide Area Information Servers, or WAIS (pronounced ways). The system permits computer users to quickly search through a huge volume of information even if it is stored at several distant locations.

The system lets users conduct searches by typing common English phrases instead of more complicated computer commands. While current systems like Dialog and Nexis require users to specify precisely the information they want, the new system can respond to a user's inferences. It initially presents a sample list of documents. The user chooses one or several, and then a "relevance feedback" program presents other documents most like the ones selected.

It will soon be possible to search through millions of items in seconds.

"This solves the problem of how to get to the information you need, getting not too much and not too little," said Esther Dyson, editor of Release 1.0, a computer industry newsletter.

This is a sharp contrast to the way services operate today, Ms. Dyson said. A computer user may need to call seven or eight separate data bases depending on the kind of information needed.

The WAIS system lets users of Apple personal computers harness a network of Thinking Machines supercomputers and smaller "server" computers to search data bases stored by Dow Jones, KPMG and several corporations and universities. Users can also read electronic mail, enter their corporate electronic libraries and summon up a wide variety of documents, newspapers and magazines.

A 'Corporate Memory'

At Thinking Machines, the WAIS system serves as a "corporate memory," allowing employees to retrieve memos, documents and other internal information. Employees who may not be working together can share expertise.

"If someone did something in Los Angeles and I'm sitting in San Francisco, I may not know about the work," said Robin Palmer, a senior manager at Peat Marwick.

WAIS delivers information over the Internet, a collection of 2,600 high-speed public and private computer networks. This Government-sponsored system of data highways is rapidly being improved and turned to commercial uses.

The market for software that allows the rapid retrieval of computerized text is small but growing, according to industry analysts. In 1989, the United States had fewer than 60,000 users; by the next year, total sales were about $120 million. The Delphi Consulting Group expects the market to grow to 160,000 users and $235 million by 1992.

"Information retrieval technology is starting to spread from supercomputers all the way down to personal computers," said Brewster Kahle, a Thinking Machines scientist who has led the WAIS experiment.

The WAIS system is built on a procedure for retrieving information developed by librarians who initially set out to computerize their card catalogues. The procedure — known in the field as Z39.50 — now has the support of the Library of Congress, Apple, Sun Microsytems Inc., Next Inc., Dow Jones and Mead Data Central.

In the future, a special directory or "white pages" will keep an up-to-date list of all the separate sources on the network.

Apple has its own electronic library project, borrowing its name, Rosebud, from the movie "Citizen Kane." The three-year-old project is based on the WAIS system, but adds features including the ability for a user to develop a personalized electronic newspaper.

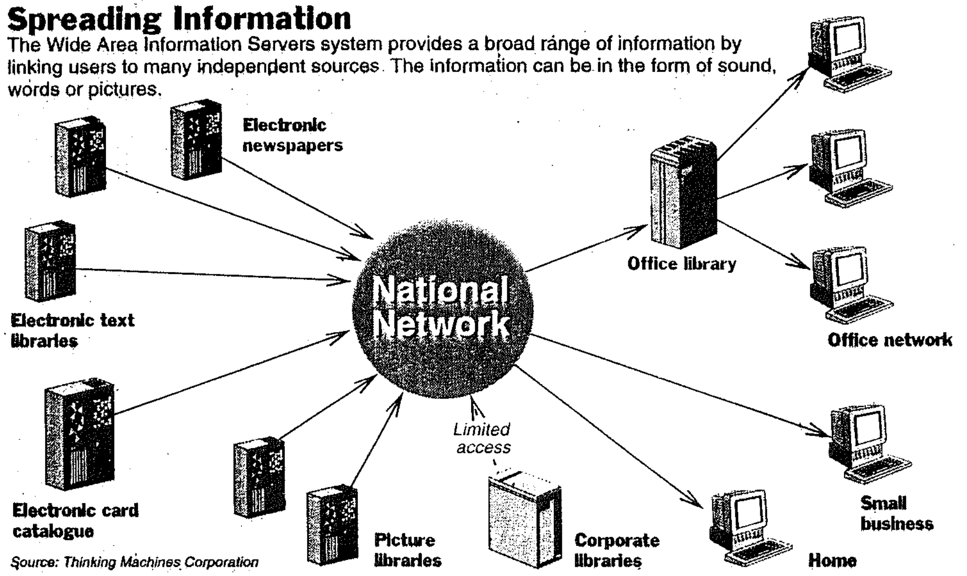
Rosebud uses special programs — called "reporters" — that let customers specify the kinds of information and news they want to retrieve from the WAIS system every day. Researchers at Apple's Advanced Technology Group said that in the future the necessary retrieval software might be a standard part of a computer's operating system.

They expect improvements in the Internet computer network to greatly lower the cost of information searches, promoting the introduction of many new services. The Government proposes to expand and improve Internet by financing a National Research and Education Network, or NREN, that could extend a high-speed computer links into schools and communities across the country.

"With things like NREN, everthing could change overnight," said Tim Oren, an Apple researcher.

Spreading Information

The Wide Area Information Servers system provides a broad range of information by linking users to many independent sources. The information can be in the form of sound, words or pictures.



Source: Thinking Machines Corporation